A Library Linked Data use case: datos.bne.es and MARiMbA

Daniel Vila-Suero

Asunción Gómez-Pérez

Faculty of Computer Science, Technical University of Madrid

Campus de Montegancedo sn, 28660 Boadilla del Monte, Madrid

http://www.oeg-upm.net

dvila, asun@fi.upm.es

Acknowledgements: B. Villazón-Terrazas, E. Montiel-Ponsoda, R. Santos, A. Manchado, M. Hernández Agustí, M. Jiménez Piano, E. Escolano.

This work is supported by the Spanish Project TIN2010-17550 for the **BabeLData** project, and by **BNE**

INRIA
Grenoble, France
9th of May 2012

Outline

- **Ontology Engineering Group**
- Library Linked Data and Motivation
- datos.bne.es project
- MARiMbA
- Results and comparison

People

Director: A. Gómez-Pérez

Research Group (30 people)

- 2 Full Professors
- 4 Associate Professors
- 2 Assistant Professor
- 5 Postdocs
- 16 PhD Students
- 1 MSc Students

Technical support (3 people)

- 2 software engineers
- 1 system administrator

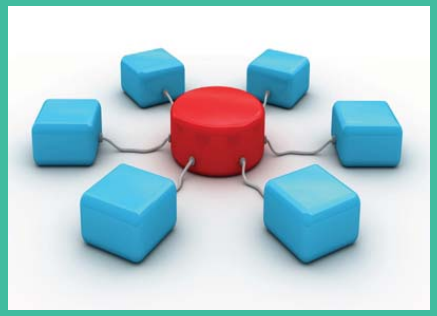
Management (3 people)

- 1 Project Manager
- 2 administrative assistants

50+ Past Collaborators

10+ visitors

Research Areas



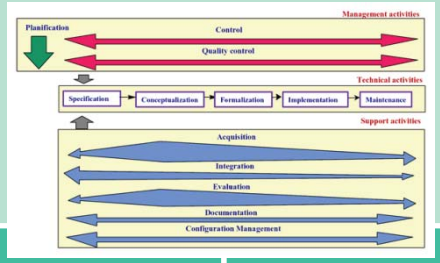
2004

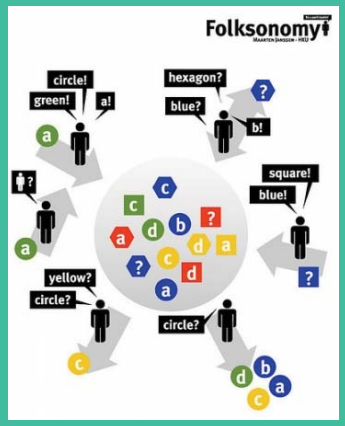
Semantic e-Science (Data Integration, Semantic Grid)

2008

Internet of Things

Ontological Engineering 1995





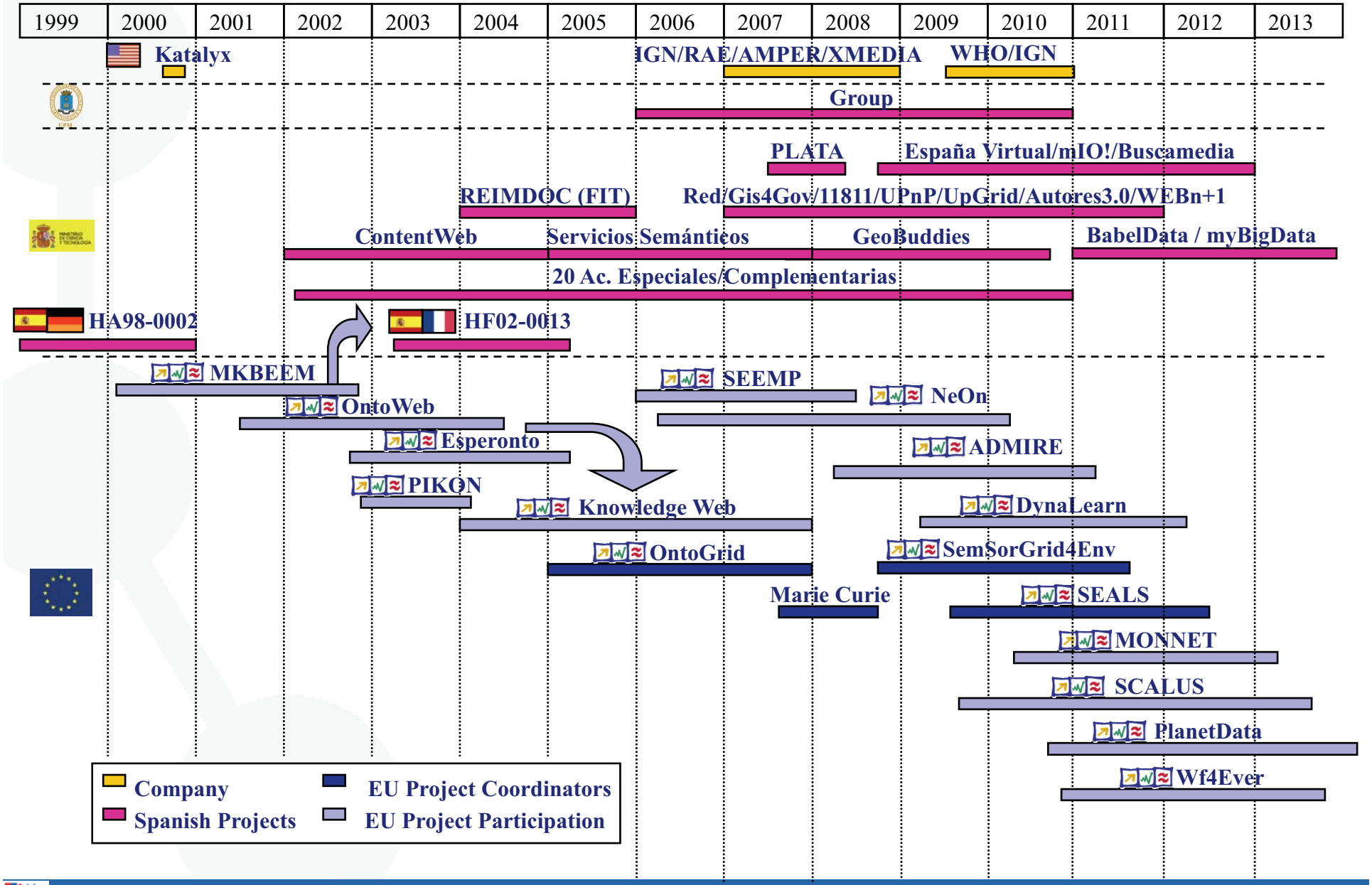
(Social) Semantic Web and Linked Data

2000

Natural Language Processing

1997

> 30 Research projects



Outline

- Ontology Engineering Group
- **Library Linked Data and Motivation**
- datos.bne.es project
- MARiMbA
- Results and comparison

Library Linked Data

- Apply Linked Data principles to library (and museums, and archives) data:

  (1) Use URIs as names for things

  (2) Use HTTP URIs so people can look up those names

  (3) Provide useful information, using the standards (RDF*..)

  (4) Include links to other URIs so that they can discover more things (not only sameAs links!)

- Growing interest from cultural institutions in the RDF data model, Linked Data, Open Data in general: *IFLA, Europeana LOD, CENL, Stanford Manifesto, W3C*.. **But why?**

W3C LLD XG introduction

- Short-lived working group: around 1 year
- "innovative ideas for specifications, guidelines, and applications that are not (or not yet) clear candidates as Web standards"
- To **help** increase global interoperability of library data on the Web, by
  - **bringing together people** from **Semantic Web,** the **library community and beyond**,
  - **identifying collaboration tracks for the future**.

W3C LLD XG results

- 3 reports: Main report, Use Cases, Vocabularies and Datasets. (http://www.w3.org/2005/Incubator/lld/)
- **Main report**:
  - **Benefits**
  - Current situation
  - Recommendation
- **Use cases report**: +50 use cases
- **Vocabularies and datasets**: Practical overview of current resources.

W3C LLD XG: Benefits

- For **users**:
  - Improved discovery and browsing of data
  - Better visibility
  - Enriched publication
- For **organizations**:
  - Bottom-up approach to data publication → more actors, different views
  - Wider choice of technologies (not only ILS vendors)
  - Lower infrastructure costs
  - Get more accessible to developer communities
  - Embrace Open Standards
- For **curators**:
  - Up-to-date directly citable by catalogers (using URIs)
  - Reduce redundancy, and duplication
  - Curators can focus on their domain of expertise (re-use)

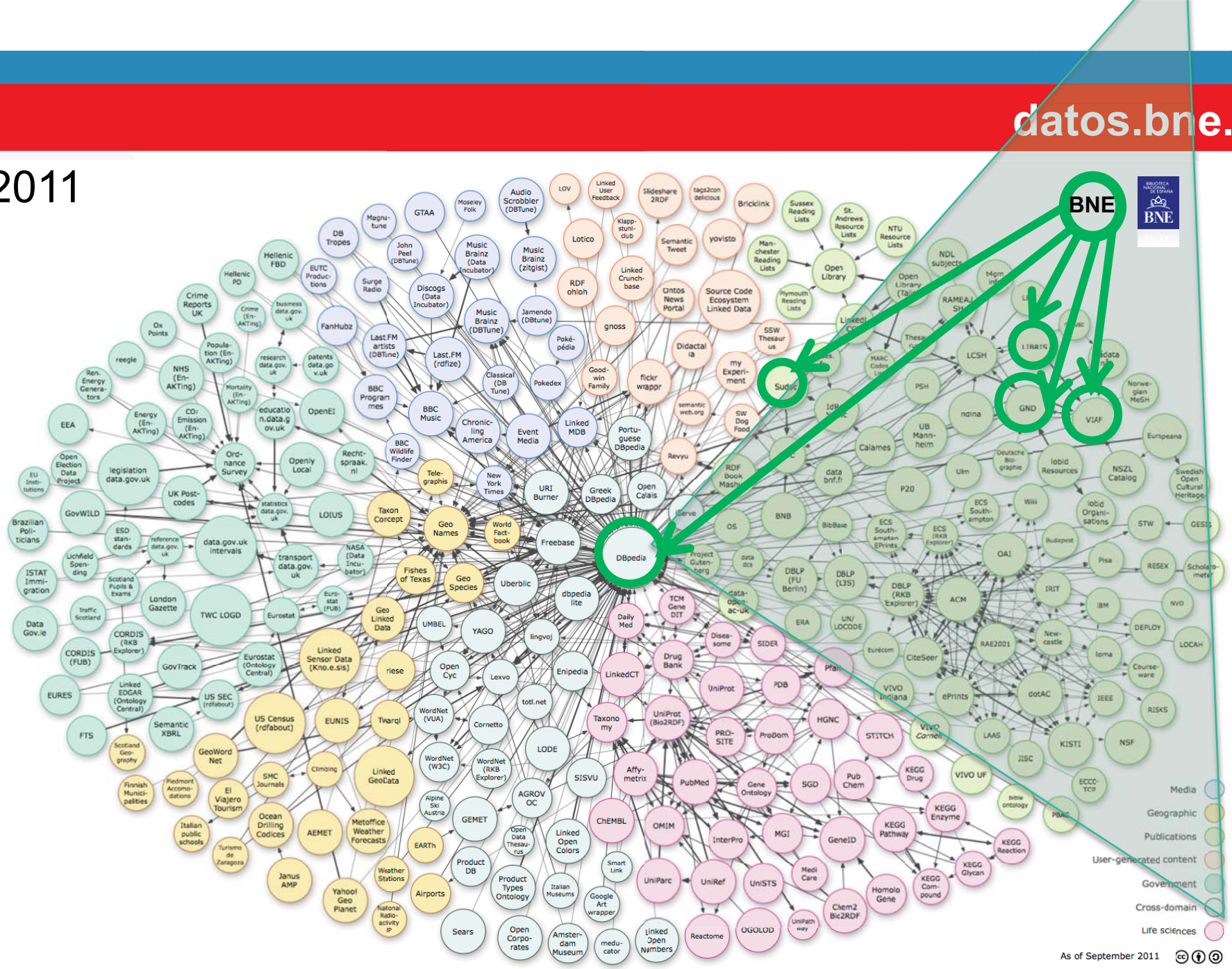
Outline

- Ontology Engineering Group
- Library Linked Data and Motivation
- **datos.bne.es project**
- MARiMbA
- Results and comparison

datos.bne.es project

- Joint project between the National Library of Spain (BNE) and Ontology Engineering Group
- Started as a small proof-of-concept project: Publishing "Cervantes" Datasets as LD
- Evolved into a bigger project:
  Publishing a significant part of the BNE catalogue
- Published in December 2011, public announcement at BNE

2011



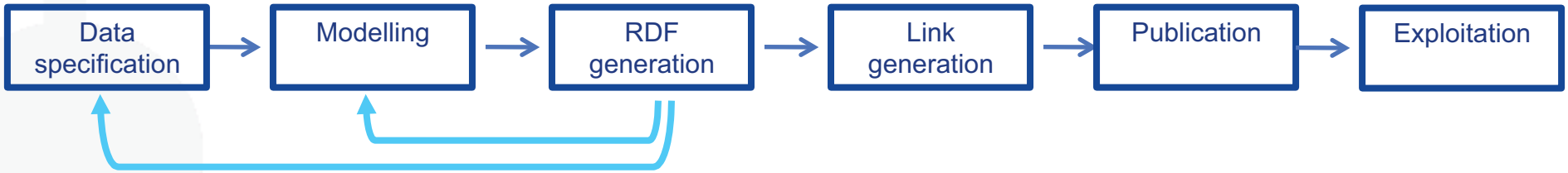
Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

datos.bne.es: Initial requirements and issues

- **Source data**: MARC 21 records, not RDB. Very flat structure difficult to map to richer models
- Domain experts (catalogers) need to be part of the mapping process.
- Data quality good but still many errors: **reporting**.
- Iterative and incremental transformation process: **measure coverage and progress.**
- Highly specialized library models: FRBR, ISBD.
- Multilinguality, collaboration with IFLA

datos.bne.es: Methodological approach

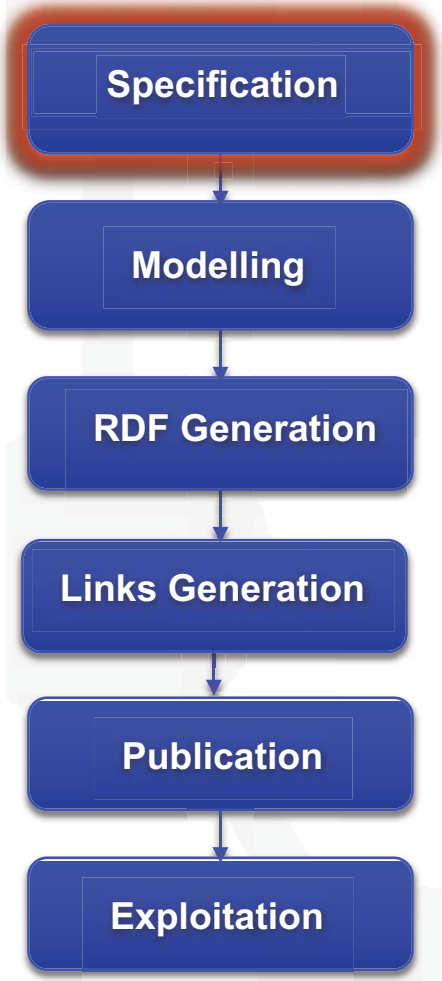
- Derived from several experiences at OEG: geolinkeddata.es, Met agency, etc. [1]



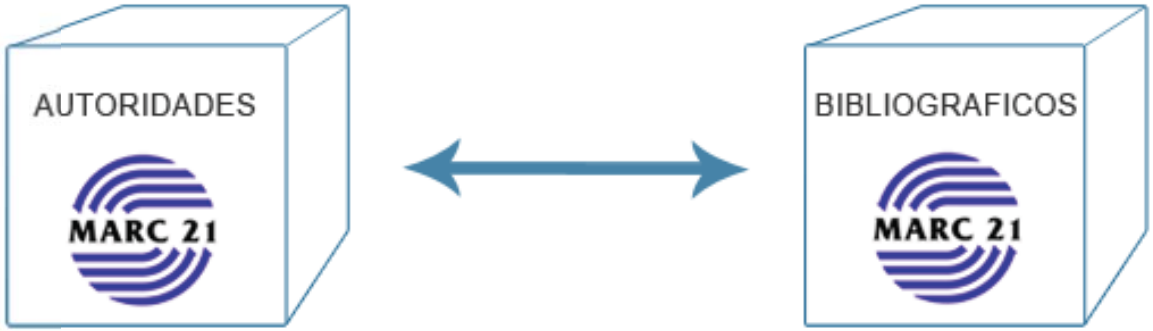
MARiMbA

- Design principle: Have more control over the different activities, allow for iterative, incremental process

[1] **Villazón-Terrazas, B**. et al., Methodological Guidelines for Publishing Government Linked Data. In D. Wood, ed. Linking Government Data. Springer.

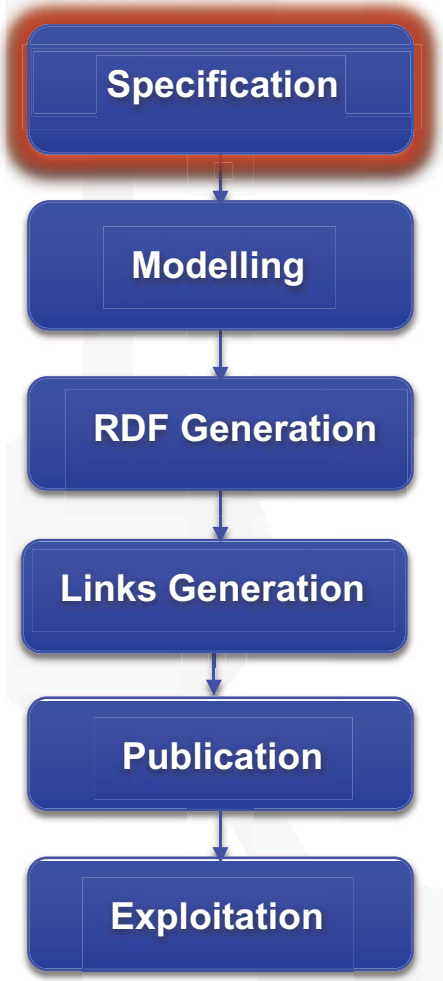


- Records in the **MARC 21 format**
- **3.9 million** bibliographical records
- **4.2 million** authority records
- Version: November, 2011



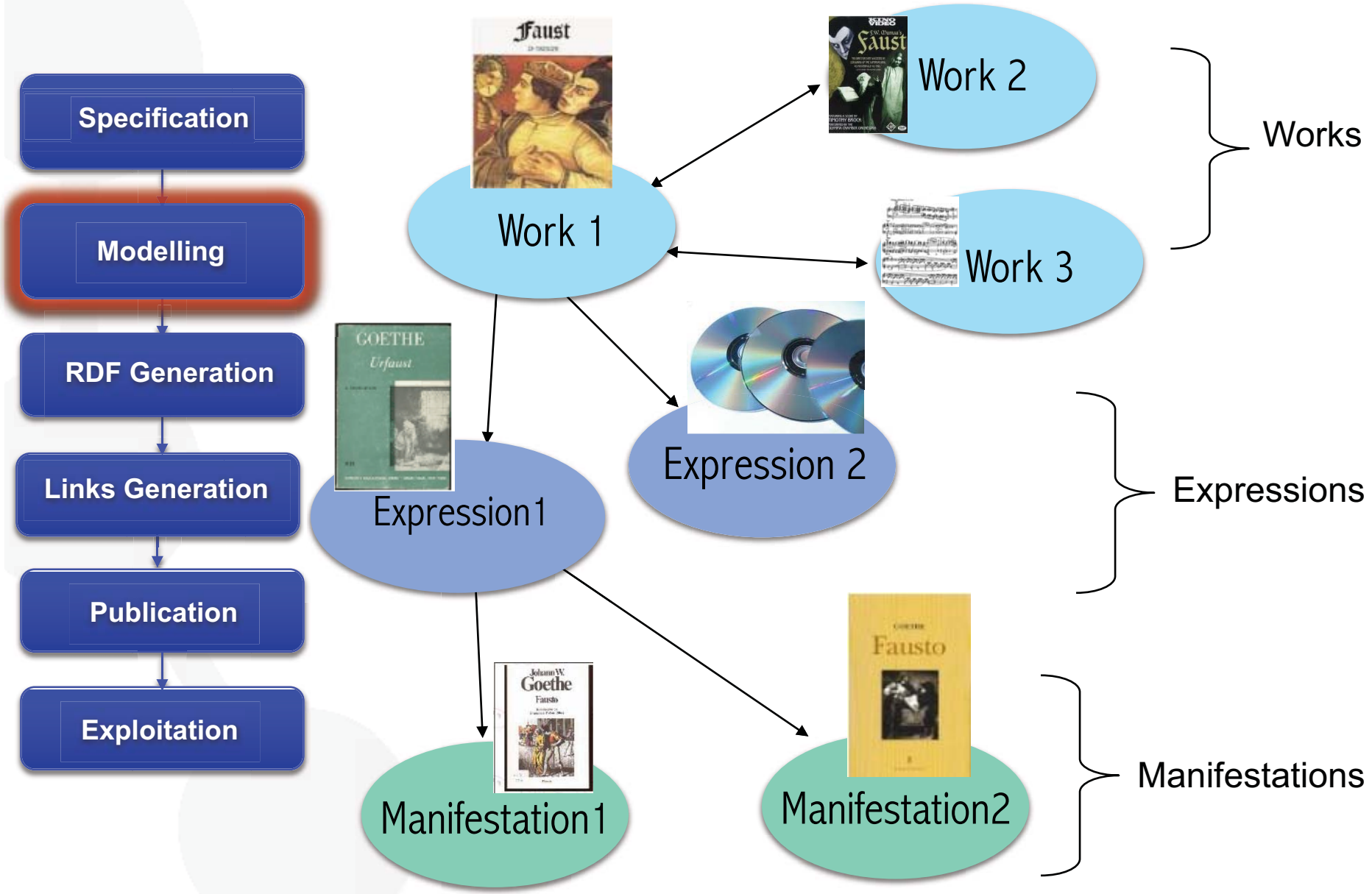
2492684	Personas
347737	Entidades
84381	Congresos
715131	Títulos
347793	Materia

76576	Mapas
320727	Grabaciones sonoras
166017	Grabados, dibujos y fotografías
35770	Manuscritos
143959	Monografías antiguas
2696560	Monografías modernas
178473	Partituras
3021	Recursos electrónicos
156634	Publicaciones seriadas
96672	Vídeos

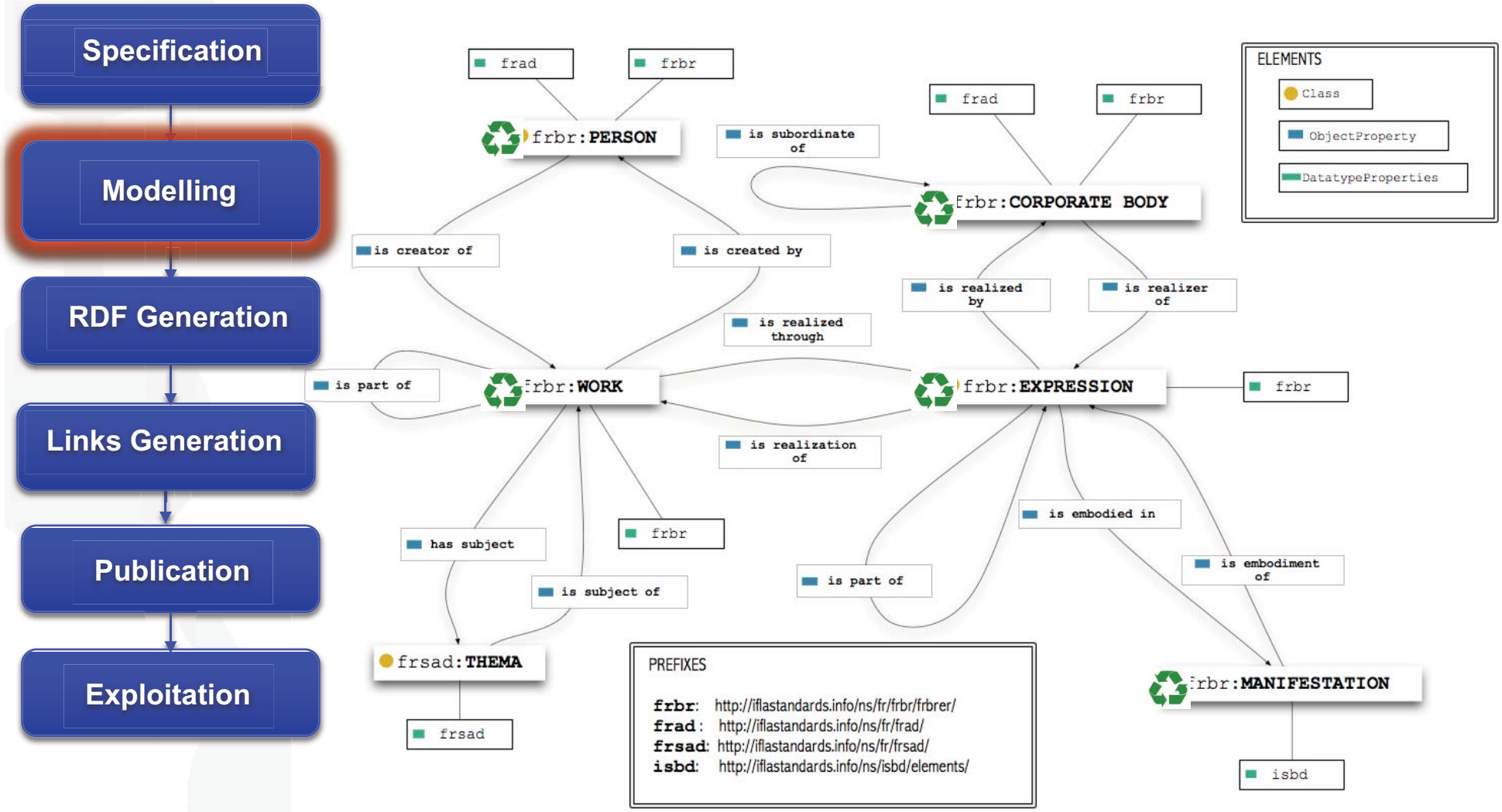


- Define here the characteristics of MARC21 records:
  - Specialized format, complex structure
  - Need for specialized tools
  - …

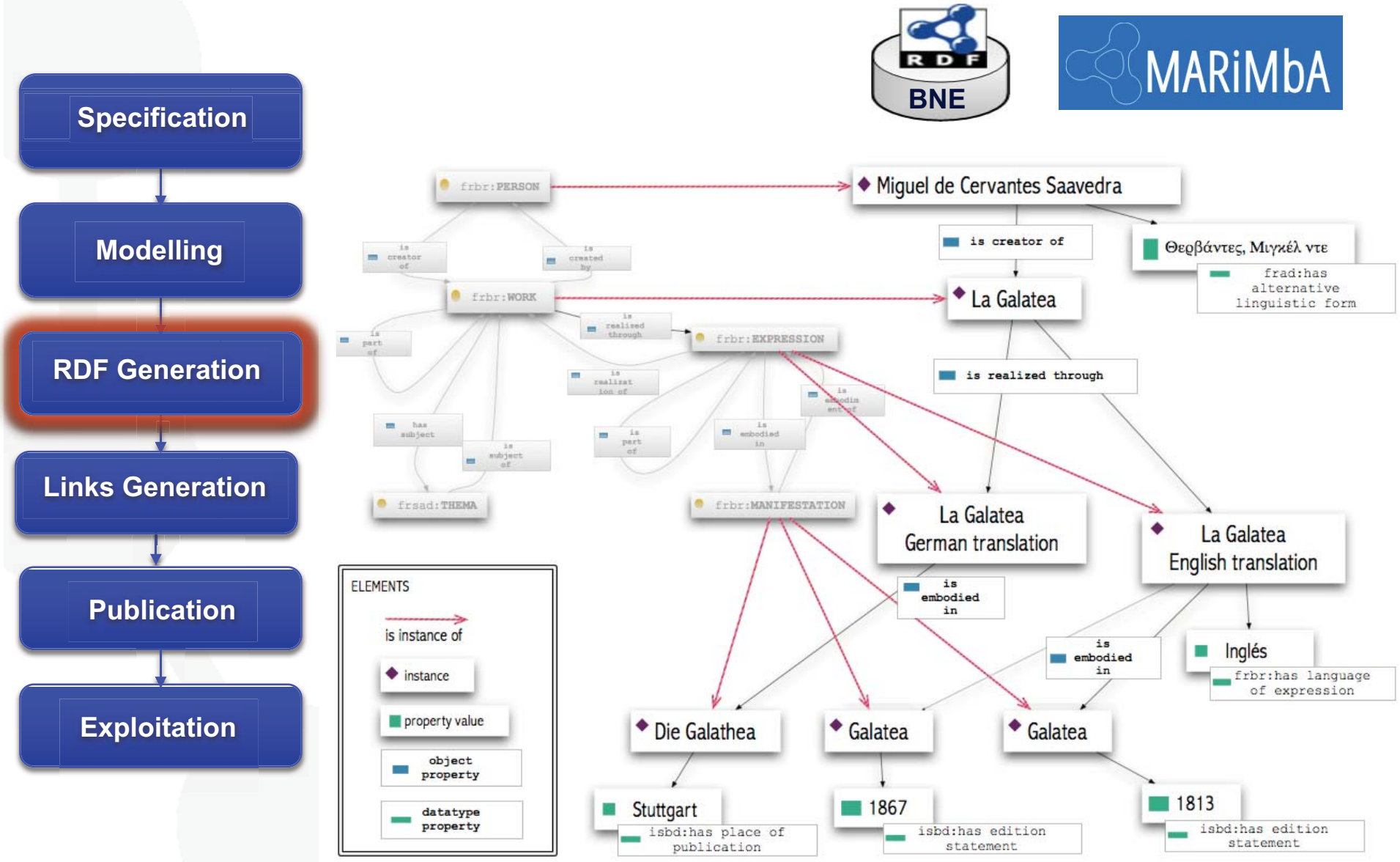
Model: FRBR at a glance



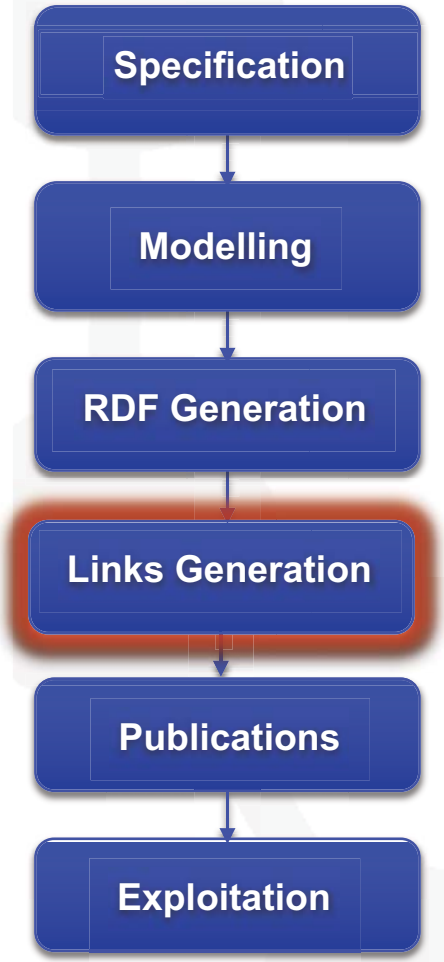
IFLA Vocabulary-based ontology

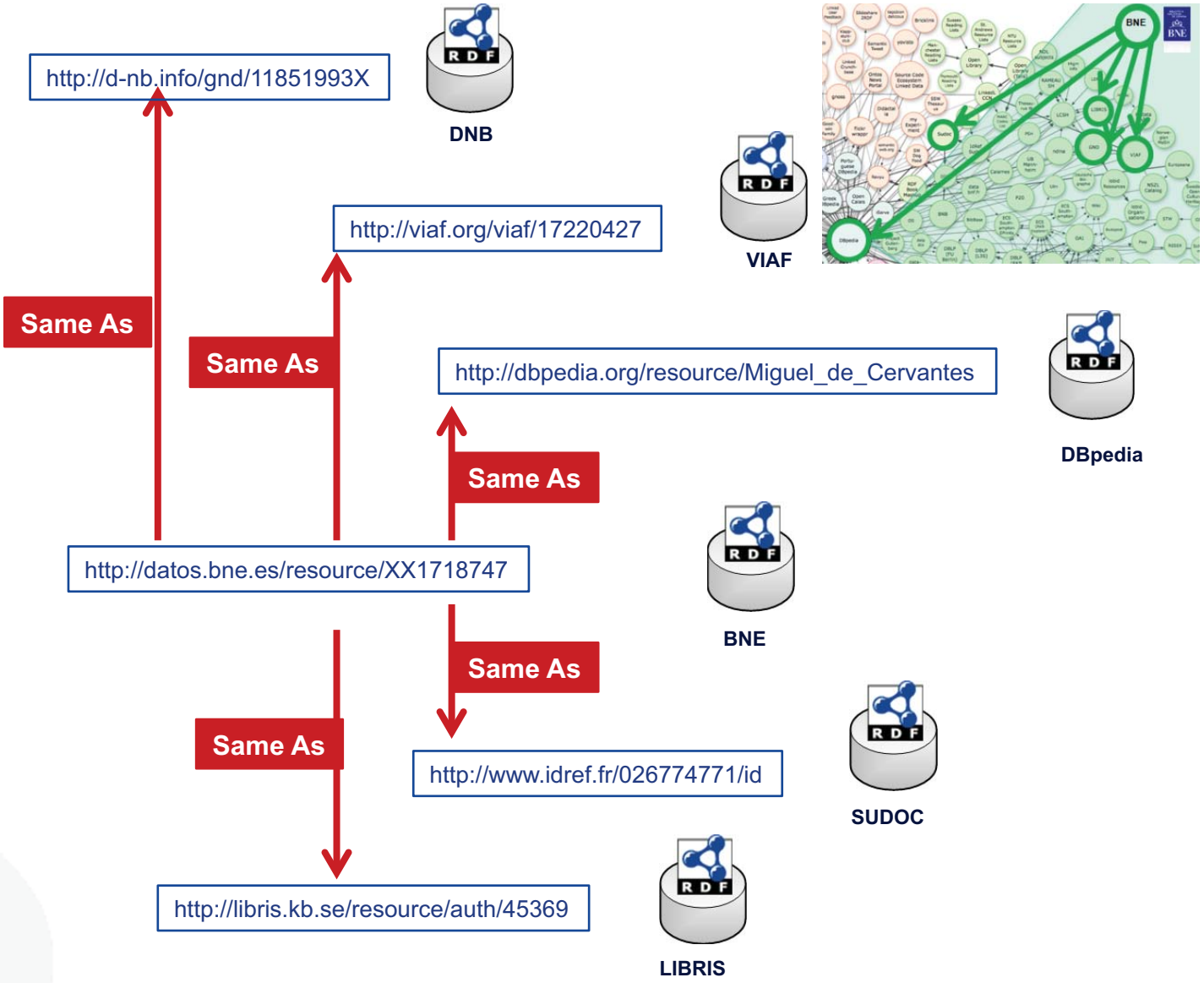


MARiMbA generates RDF using RDFS/OWL ontologies

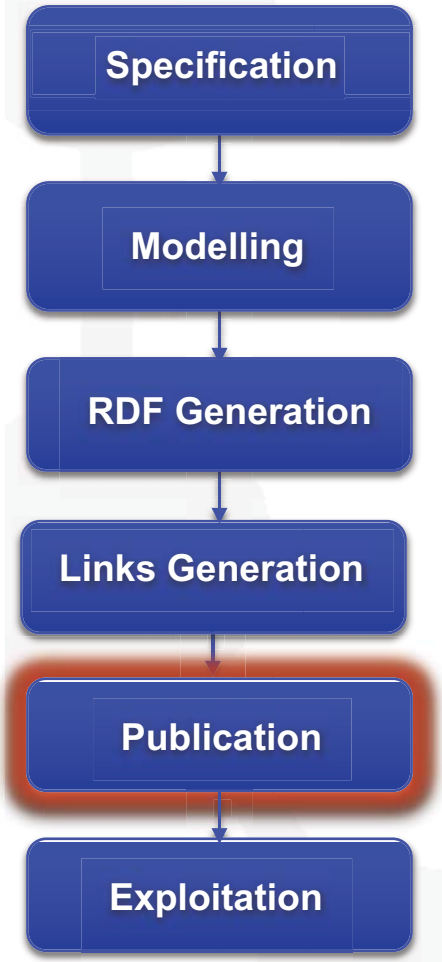


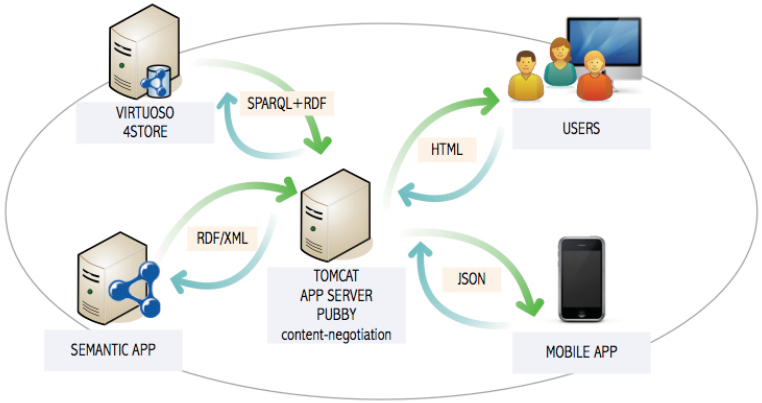
MARiMbA links with other resources:
VIAF, DNB, SUDOC, LIBRIS, DBpedia





Publication



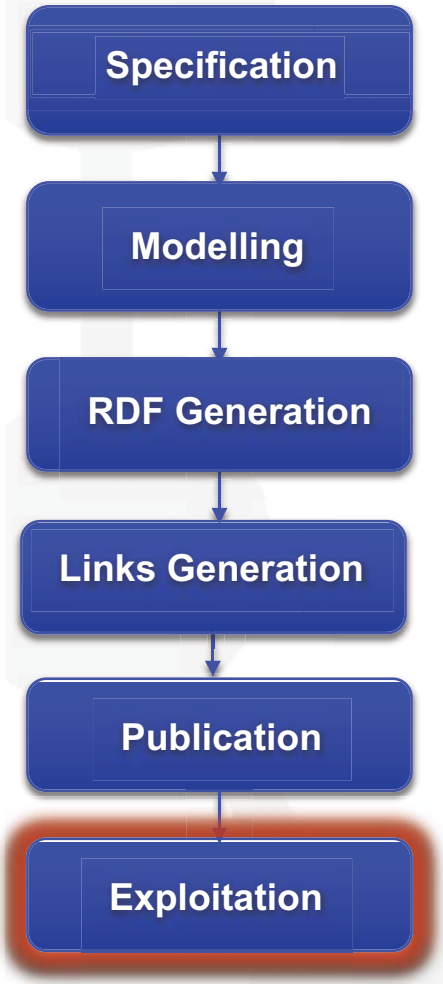


Data publication

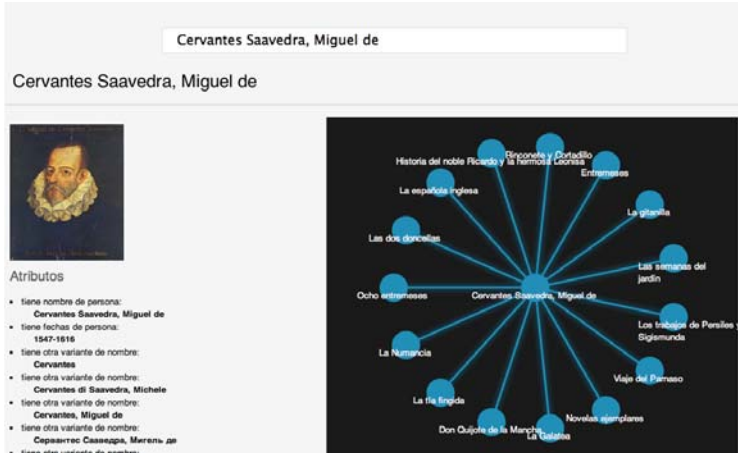
Metadata publicacion using VOID

To facilitate the discovery

- Register in CKAN your dataset
- Use **sitemap4rdf** to generate the site map
- Upload the site map to **Google** and **Sindice**



Web Interface



http://linkeddata3.dia.fi.upm.es/bne-demo

SPARQL Queries:

http://datos.bne.es/sparql

```
select distinct COUNT(?Obras) where {

http://datos.bne.es/resource/XX1718747

 <http://iflastandards.info/ns/fr/frbr/frbrer/P2010>
?Obras
```

URI Cervantes

Is author

Outline

- Ontology Engineering Group
- Library Linked Data
- W3C Library Linked Data Incubator Group
- datos.bne.es project
- **MARiMbA**
- Results and comparison

MARiMbA

- "A MARC Mappings and RDF generator"
- Supports the ETL process by:
  - Analysing the source records.
  - Generating mapping templates (spreadsheets) based on the analysis, providing useful information to users (domain experts)
  - Transforming MARC records to RDF.
  - Providing a light-weight SPARQL endpoint to query/browse the resulting RDF (using FUSEKI).
- Three step process:

1. Analyse records and generate mapping templates

2. Assign mappings using mapping templates

3. Generate RDF and produce a report

MARC21

- Machine-readable format widely used for representation and exchange

- Different communication formats:
  - **MARC 21 format for Bibliographic Data**
  - **MARC 21 format for Authority Data**
  - Others: Holdings, Classification, etc.

- Three main elements:
  - **Record structure**: ISO 2709. Fields, indicators, subfields…
  - **Content designation**: "Meaning" of codes and conventions
  - **Content**: Defined outside the MARC standard (ISBD, AACR..)

MARC21 record structure

- Authority record: Camus, Albert*

Control Field	
001	XX1721208
005	200012181124
008	901120nn aijnnaabn n aaa
016	$a BNE19900178994
040	$a SpMaBN $b spa $c SpMaBN $e rdc $f embne

Field	Subfield	Content				
100	10	$a	Camus, Albert			HEADING 1XX
		$d	1913-1960			

670	$a El mite de Sísif, 1987 $b port. (Albert Camus)	
670	$a Dic. de filosofía, de J. Ferrater Mora, 1980$b(Camus., Albert (1913-1960); n. Mondovi, Argel)	
670	$a Aut. BN-OPALE, 1995 $b (Camus, Albert)	

* http://datos.bne.es/resource/XX1721208

MARC21 record content designation

- Authority record: Camus, Albert*

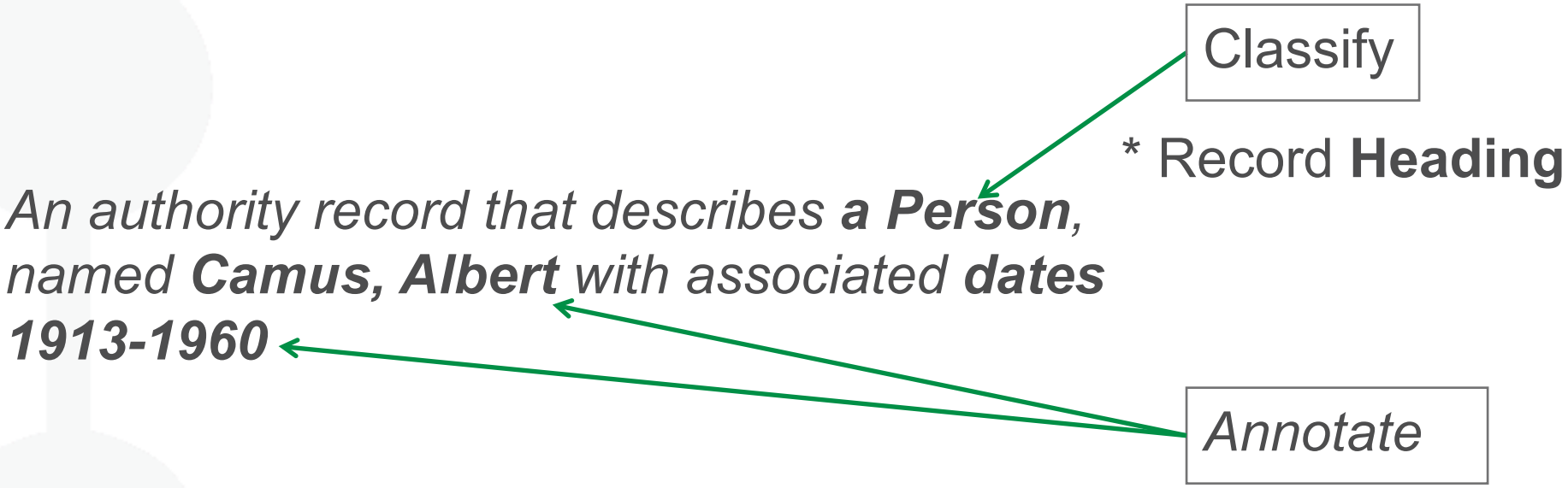
Control Number		001	XX1721208		
					HEADING – Personal Name
Personal name	**Name**	100	10	$a Camus, Albert	**100**
Dates associated with name				$d 1913-1960	

- Human reading:

  *An authority record that describes **a Person**, named **Camus, Albert** with associated **dates 1913-1960***

* http://datos.bne.es/resource/XX1721208

Mapping process intuitively



MARC 21 record (Input)	Action	RDF (Output)
100 $a $d	Classify	rdf:type foaf:Person
100 $a Camus, Albert	Annotate	foaf:name "Camus, Albert"
100 $d 1913-1960	Annotate	frbr:P3040 "1913-1960"

Mapping process more in detail

- **Classify**: Exploiting the heading field and subfield **codes**.

  100 $a $d → Person (it has a personal name)
  100 $a $d $t → Work (it has a title)

- **Annotate**: Using subfield codes and the **content**.

  100 $a "Camus, Albert" → foaf:name "Camus, Albert"
  100 $t "La Peste" → frbr:workTitle "La Peste"

- **But,** what about the **relationships** between the entities?

  The work "La Peste" was created by Albert Camus

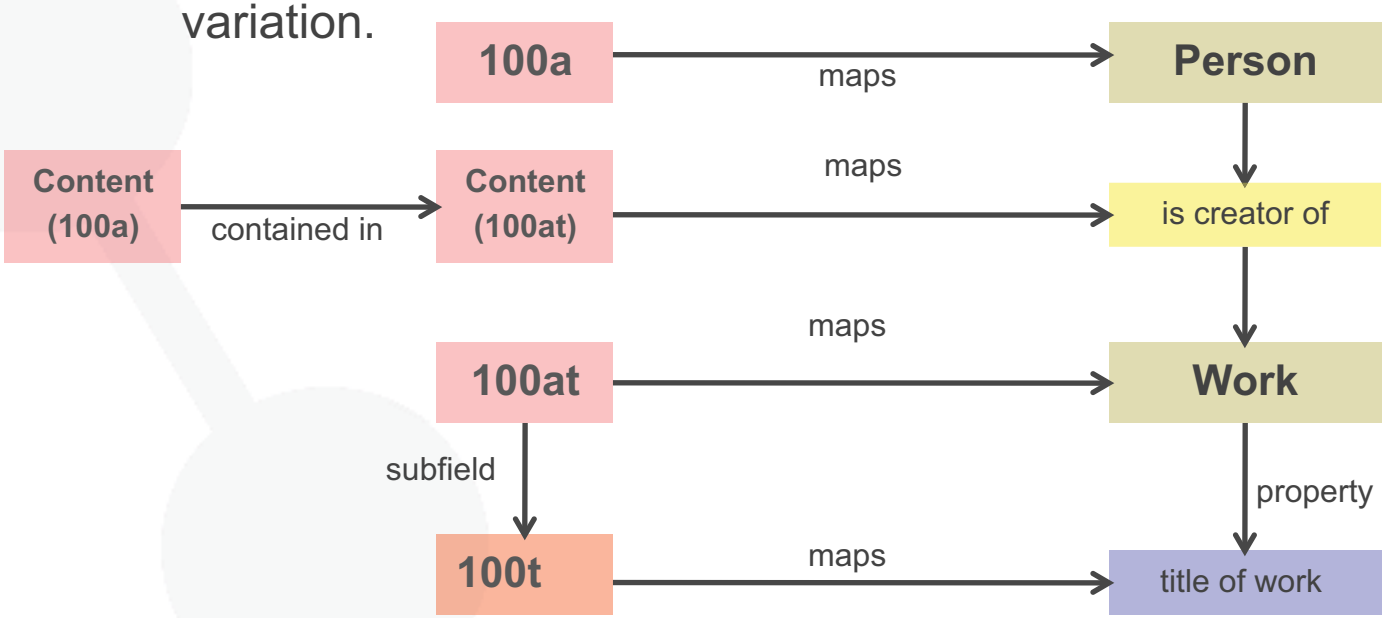
Let's see an example

Mapping process more in detail (to be refined)

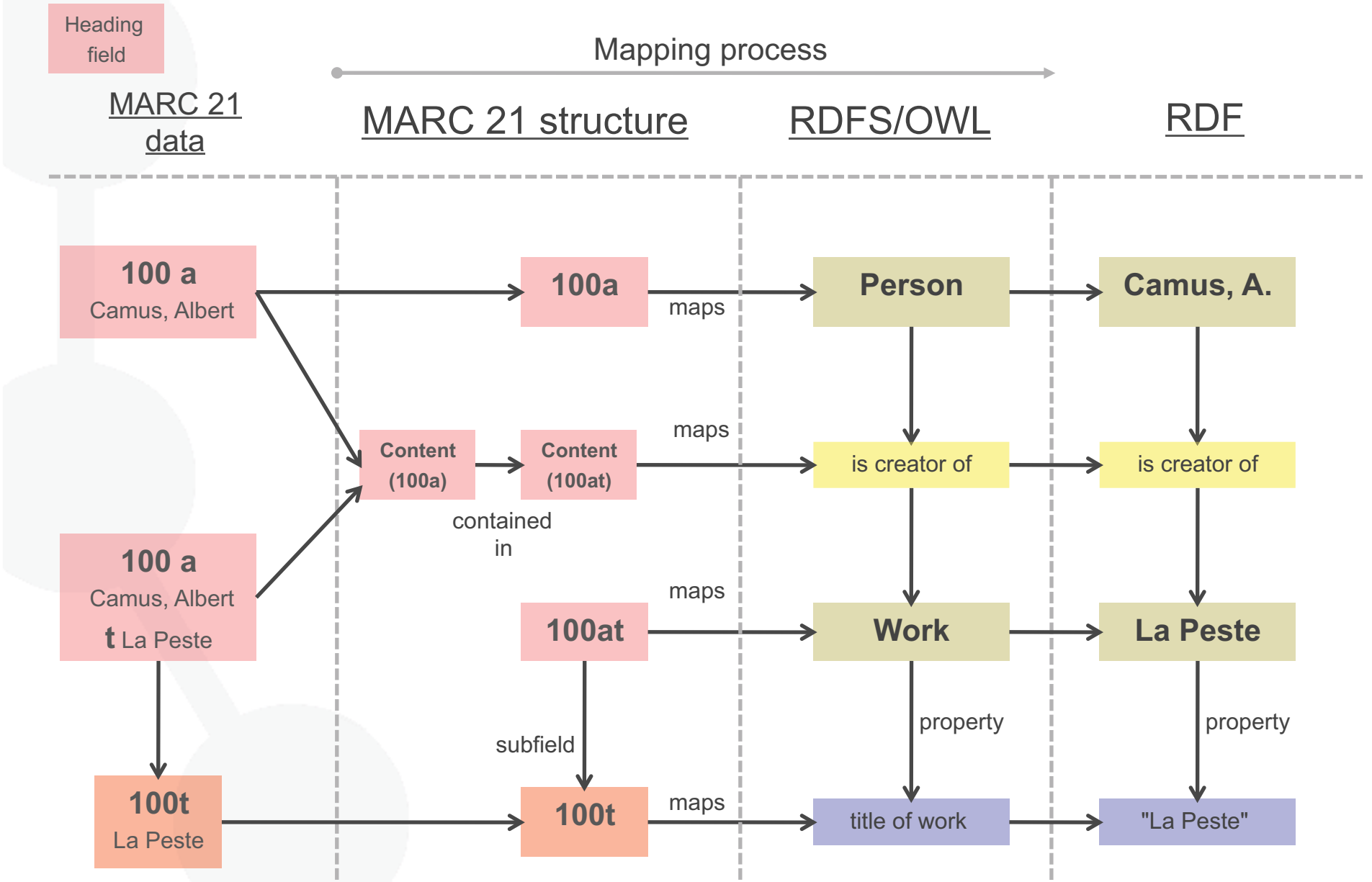
- Similar to mapping ontologies, but:
  - **Classes**: Are defined in terms of the MARC heading field and subfield codes

    100atl → Expression ; 110a → Corporate Body
  - **Properties**: Are defined in terms of field+subfield codes

    100a → name ; 100t → title of work
  - **Object properties**: Are defined in terms of heading content containment + variation.



Mapping full example with record and instance data



Mapping process more in detail

- Relationships between records are **not explicit** in MARC.

Goal: The work "La Peste" was created by Albert Camus

R1: Camus, Albert record

001 XX1721208

100 10 $a Camus, Albert $d 1913-1960

Common

Person

R2: La Peste record*

001 XX1910518

100 10 $a Camus, Albert$d1913-1960 $tLa peste

Common **Diff**

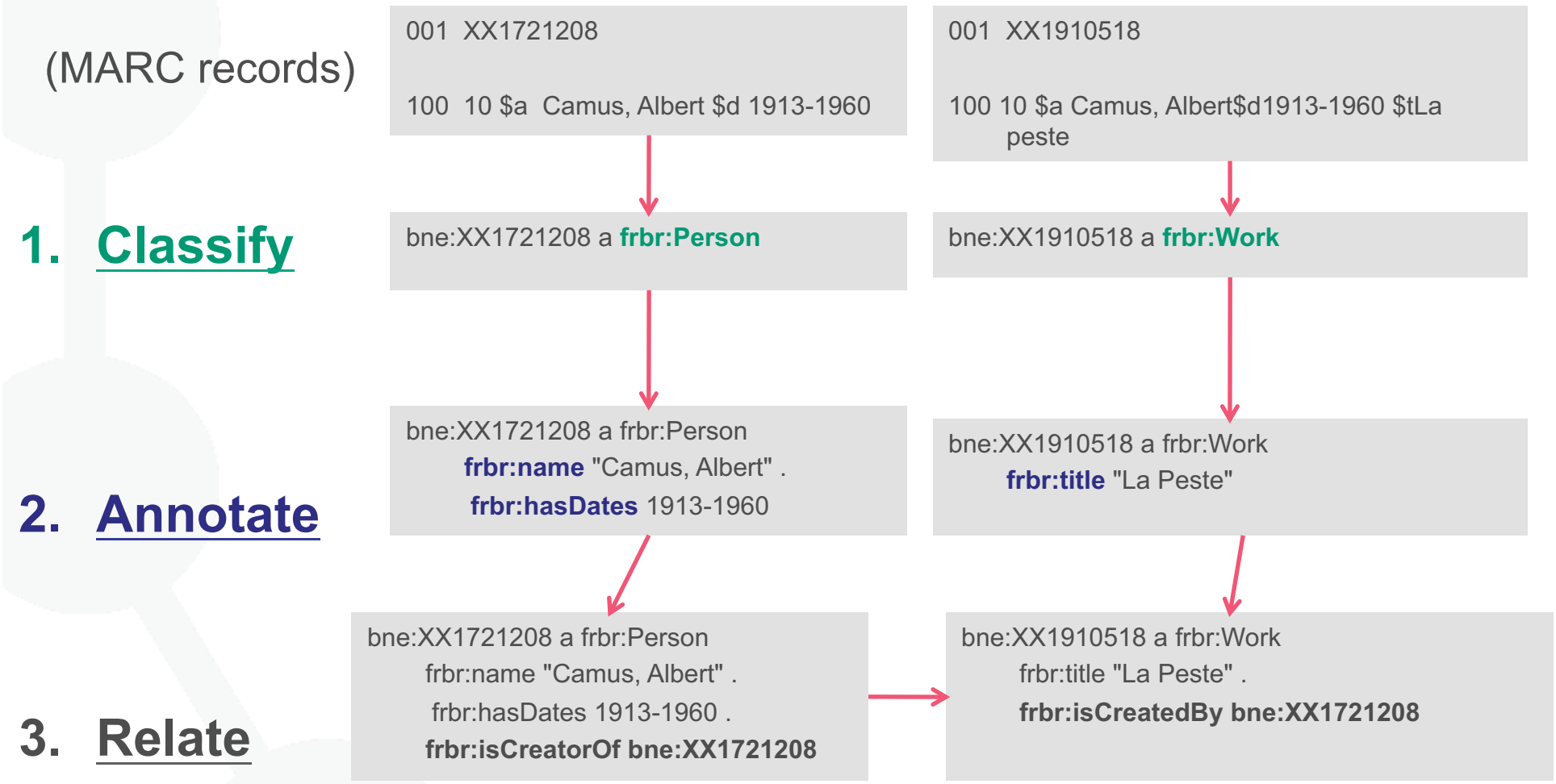
Work

We know the type of R1 and R2, and we look at the heading diff

bne:XX1721208 **frbr:isCreatorOf** bne:XX1910518

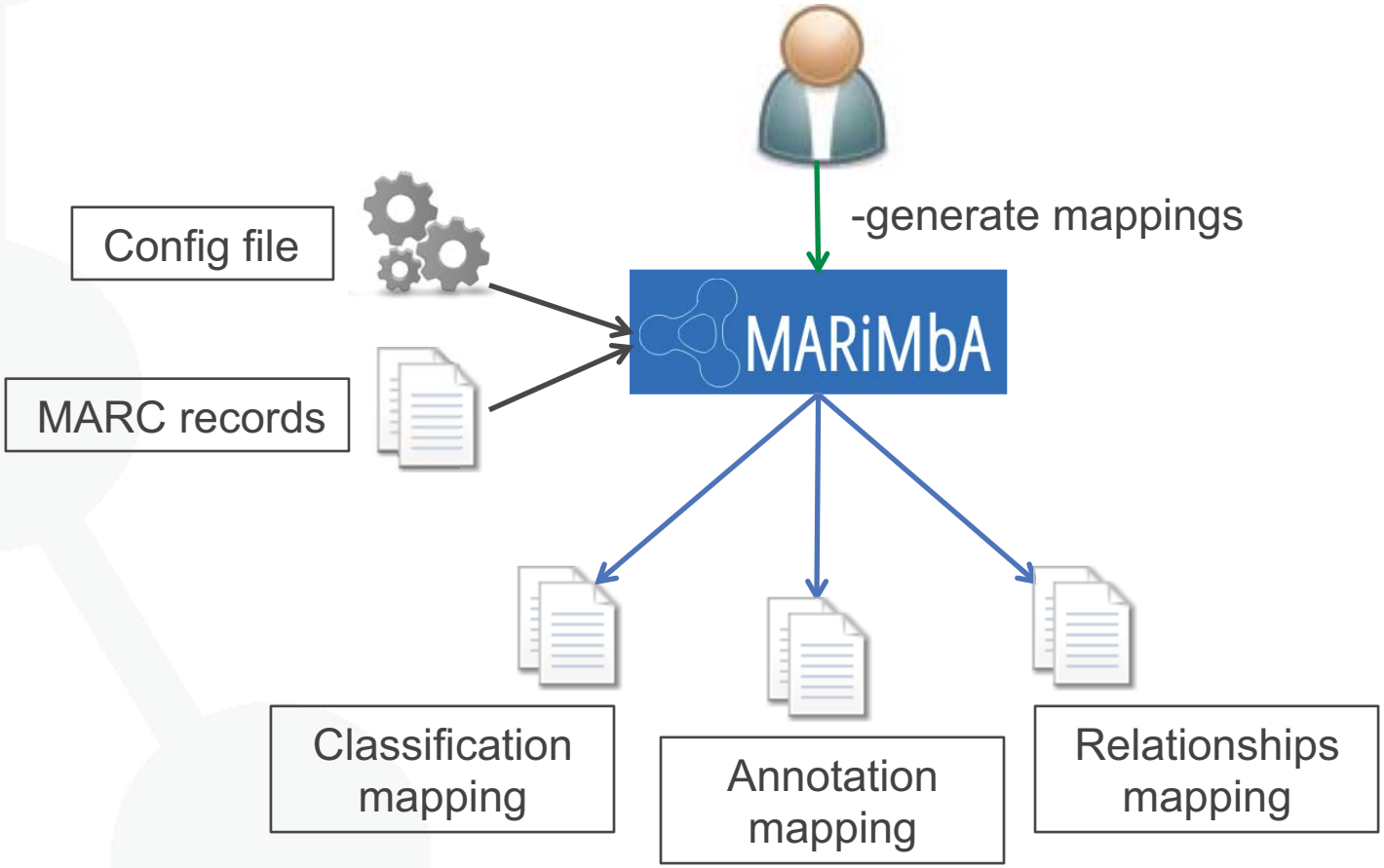
* http://datos.bne.es/resource/XX1910518

Mapping process summary



MARiMbA step 1: Analysis and template generation

- 3 steps of mapping **Classify**, **Annotate**, **Relate**
  → 3 CSV templates based on the source data



MARiMbA step 2: Assign mappings

- Three spreadsheets:

	A	B	C	
1	Combination	Records count	Example	Mapping value
2	a	881115	[$aYcardo, Maruja]	Person
3	at	685861	[$aGual, Adri‡, $t Donzell qui cerca muller]	Work

Classification mapping

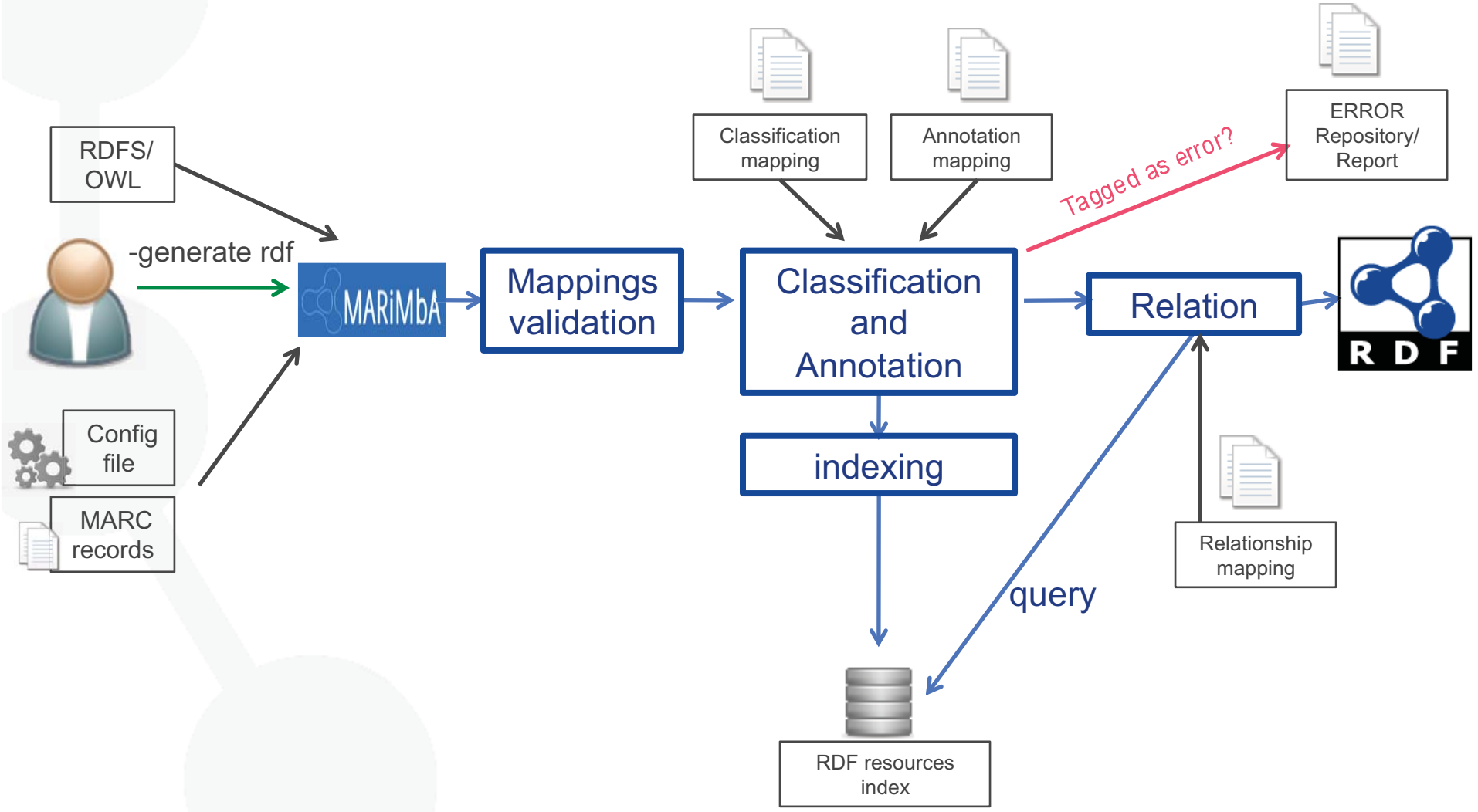
	A	B	C	Mapping value
1	Combinatio	Records c	Example	
2	260b	3892943	Imp. Lemercier et C.ie	http://iflastandards.info/ns/isbd/elements/P1117
3	245a	3871272	Polka elegante para piano	http://iflastandards.info/ns/isbd/elements/P1004
4	040a	3870286	SpMaBN	
5	260a	3845589	[Paris	http://iflastandards.info/ns/isbd/elements/P1016
6	300a	3827202	2 p.	http://iflastandards.info/ns/isbd/elements/P1022
7	245h	3751432	M´sica impresa	http://iflastandards.info/ns/isbd/elements/P1185
8	260c	3714336	1865-1869]	http://iflastandards.info/ns/isbd/elements/P1018
9	300c	3526395	22 cm	http://iflastandards.info/ns/isbd/elements/P1024
10	016a	3103593	bipaBNE20070006028	

Annotation mapping

	A	B	C	D
1	Variation	Sample	Relationship	
2	n	Sociedad Didáctico-Musical Método especial de solfeo cor	frbr:P2057; frbr:P2058	
3	o	Bach, Johann Sebastian 1685-1750 Partita laúd		
4	p	Iglesia Católica Synodus Episcoporum (2008) Verbum Dor	frbr:P2057; frbr:P2058	
5	q	Isabel María Isabel María López Campos The face		
6	r	Chopin, Fryderyk 1810-1849 Valses piano mi menor		
7	s	Beethoven, Ludwig van 1770-1827 Fantasien pia		
8	t	Castellà Andreu, Josep M. La función constituci		

Relationships mapping

MARiMbA step 3: RDF generation process overview



Open (Research) questions

- Areas for effective automation:
  - **Classification** phase: Learning algorithms seem good candidates (we have well curated training data).
  - **Relate** phase: Blocking strategies, string similarity metrics
  - **Metadata content granularity**: Can we derive mapping rules directly from models (e.g. ISBD) or cataloguing rules (e.g. AACR)?
- Curation workflow/feedback:
  - Can we define a protocol for continuous improvement of data through the ETL process? Metrics? QA?
  - Can mapping rules and cataloguing rules be used to automatically validate resources?
- Update process:
  - Protocol for incremental updates, changes propagation.
- Linking to external resources: techniques for cross-lingual instance matching

Outline

- Ontology Engineering Group
- Library Linked Data
- W3C Library Linked Data Incubator Group
- datos.bne.es project
- MARiMbA
- **Results and comparison**

Results: datos.bne.es

- Total number of authority records: **4.100.000**
- Total number of bibliographical records: **2.390.140**
- Total number of RDF triples: **58.053.215**
- Number of links: (15% authorities): **587.520**
- Linked sources:
  - VIAF
  - SUDOC (French collective university catalogue) FR
  - GND (German National Library of authorities) GER
  - LIBRIS Sweden
  - DBPedia
  - Soon BNF

Tools comparison

Feature	Metamorph (DNB)	Marc2rdf (NO)	MARiMbA (BNE)
Users	API, technical users	YAML mapping language	**Librarians Catalogers**
Formats	Authority, Bibliographic	Bibliographic	Authority, Bibliographic
Encodings	MARC, PICA+	ISO	ISO, MARCXML
Granularity	Record content designation	Content transformation	Record content designation
Source data coverage	Not controlled	Not controlled	**Covers all possibilities through analytic. process**
Error reporting	NO	NO	Limited
Degree of Automation	Limited	Limited	Limited
Complex linking	NO	NO	Yes

Datasets comparison (initial review)

Feature	DNB	BNF	BNE
Data	Authority, Bibliographic	-Authority +Bibliographic	Authority, Bibliographic
Source Data	MARC, PICA+	MARC	MARC
Granularity	++	++	+
Source data coverage	unknown	unknown	**High**
Degree of links between resources	Low, almost flat resources	Medium	High
Update	Automatic	Unknown	Bulk transformation
Complex linking	NO	NO	Yes

Questions

Thank you very much!

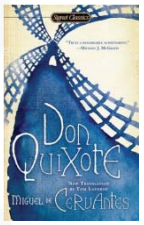
Questions and comments are very welcomed

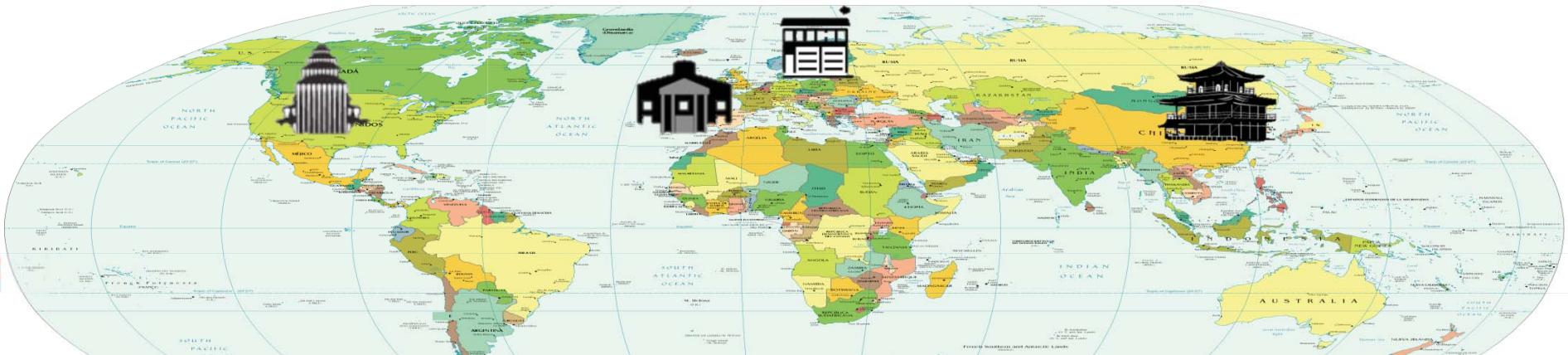
Email: dvila@fi.upm.es

Acknowledgements

Who had translated “Quixote” to other languages?

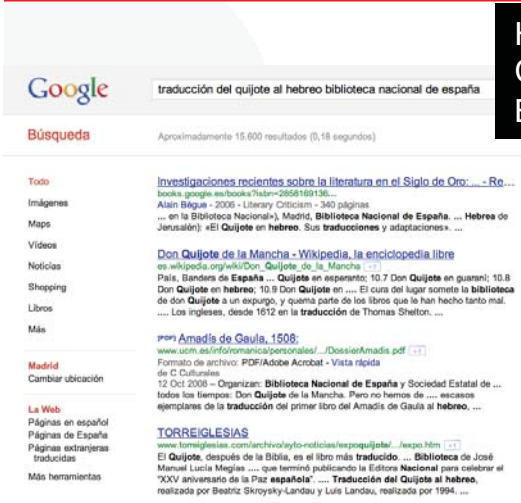
- Multiple sources of multilingual data
- The local information may be incomplete
- The remote information is not available





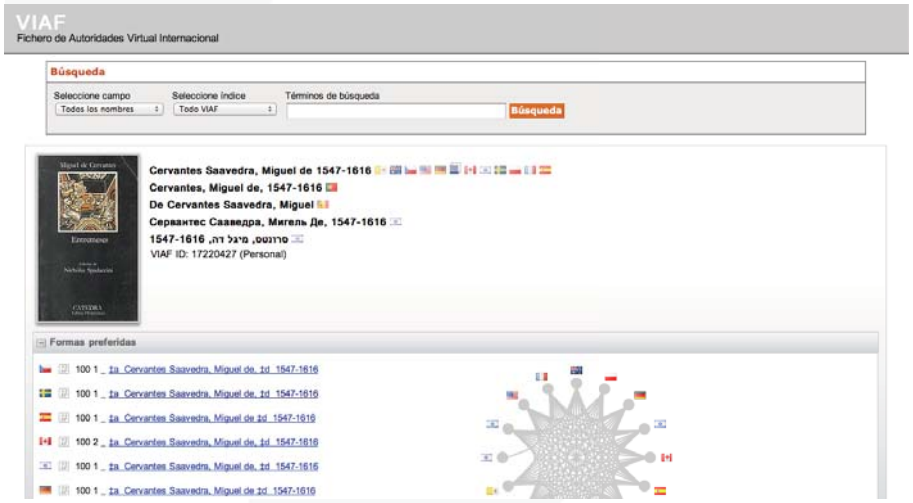
Complex queries on different data sources

How many works written by Miguel de Cervantes are registered to the BNE and to the DNB?



http://www.bne.es/

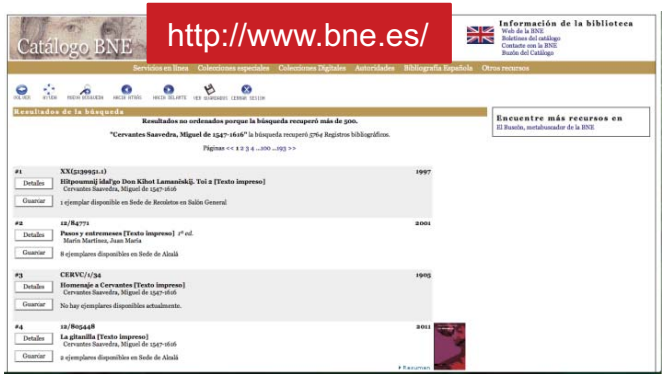
http://www.viaf.org/



http://d-nb.info

Data from different libraries exposed via Web

BNE Database

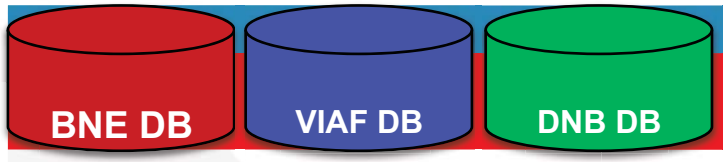


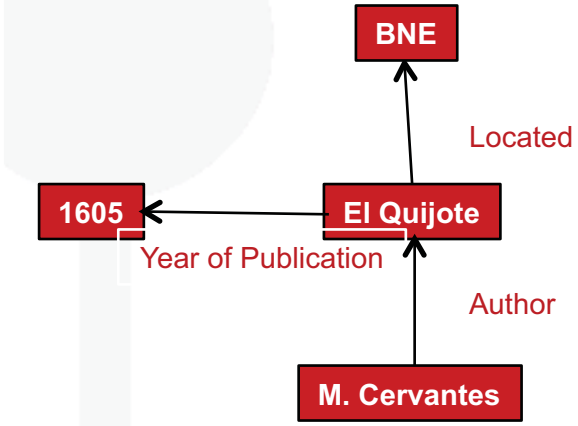
How many works written by Miguel de Cervantes are registered to the BNE and to the DNB?

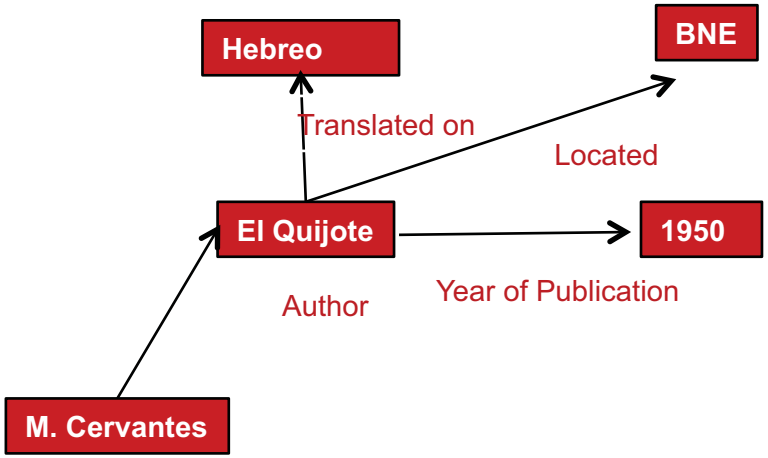
VIAF Database

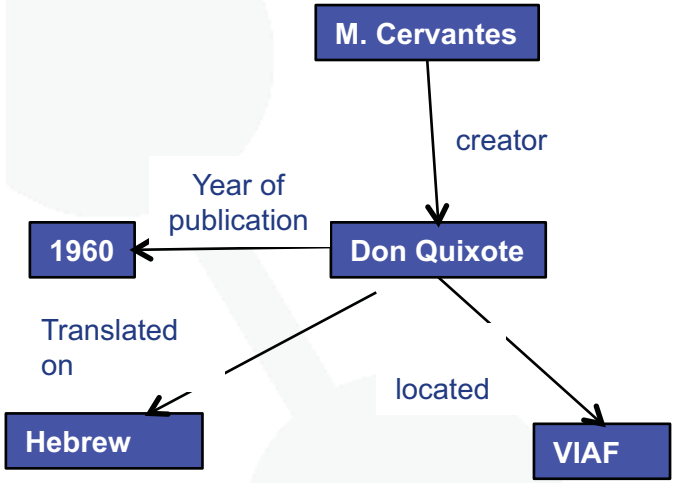
DNB Database

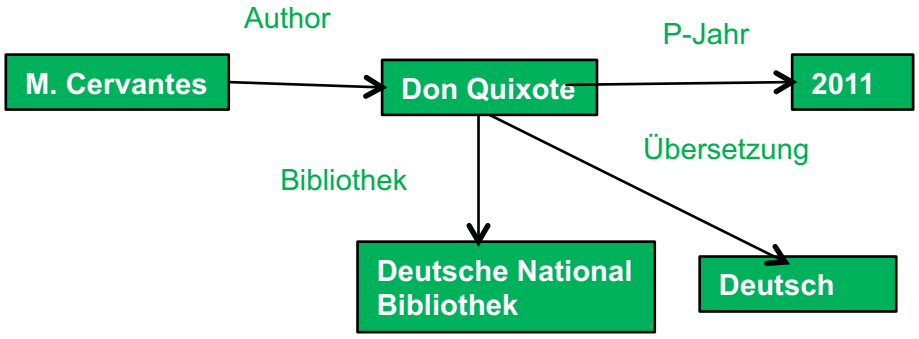
Data integration



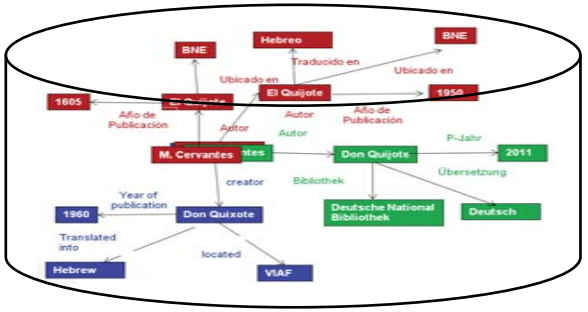


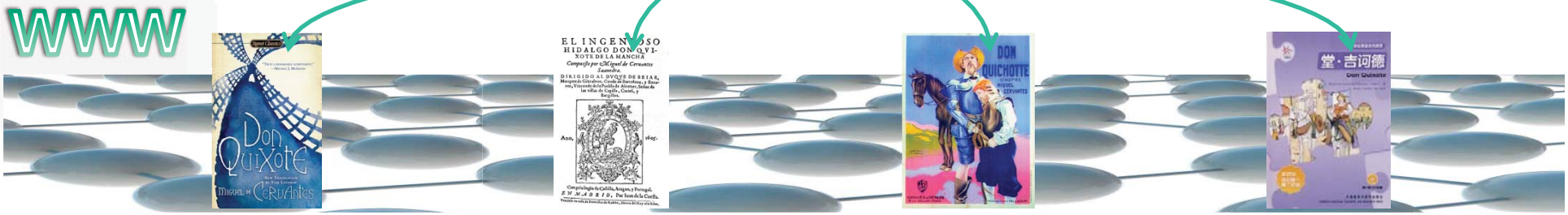


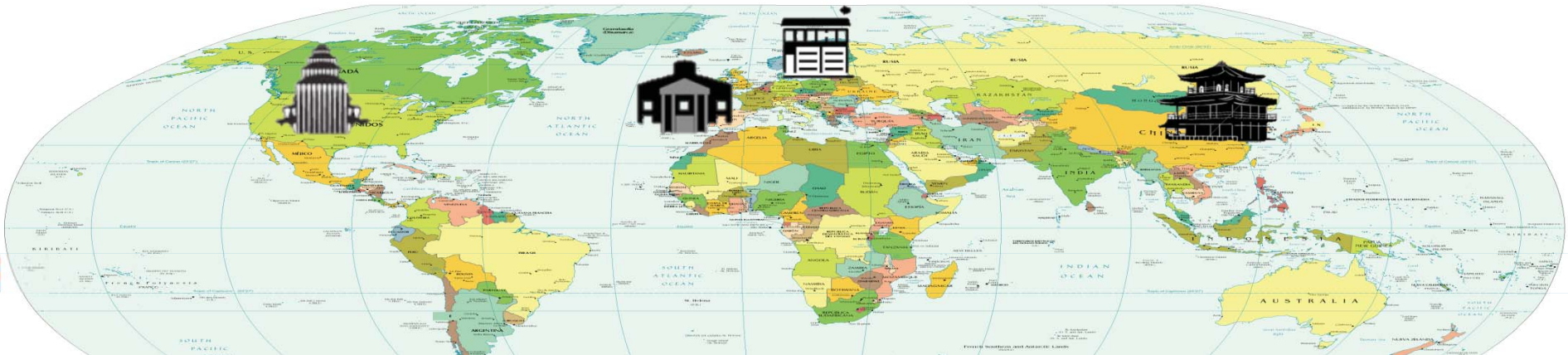




What do we need in order to answer this question?





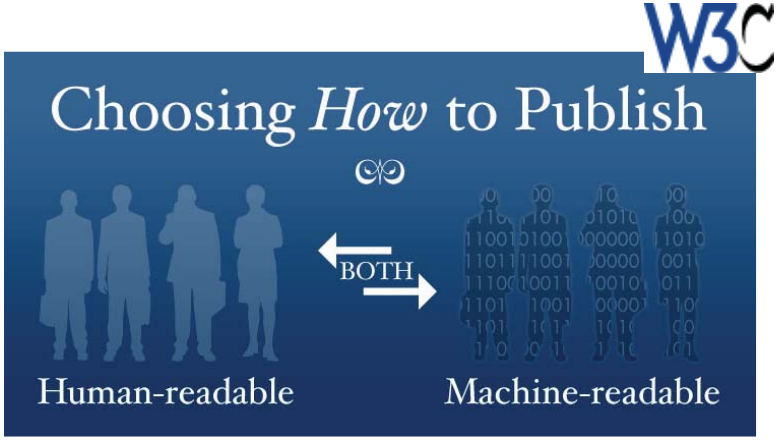


Content

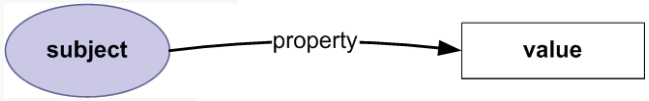
1. The concept from an intuitive manner
2. Fundamentals
3. The process
4. Marimba
5. Demo
6. Conclusions

Usefulness of Linked data

- To combine data
  - From heterogeneous sources
  - In different formats
  - With different levels of detail
  - In different languages
  - From different countries
- To facilitate data integration



RDF(S) Models



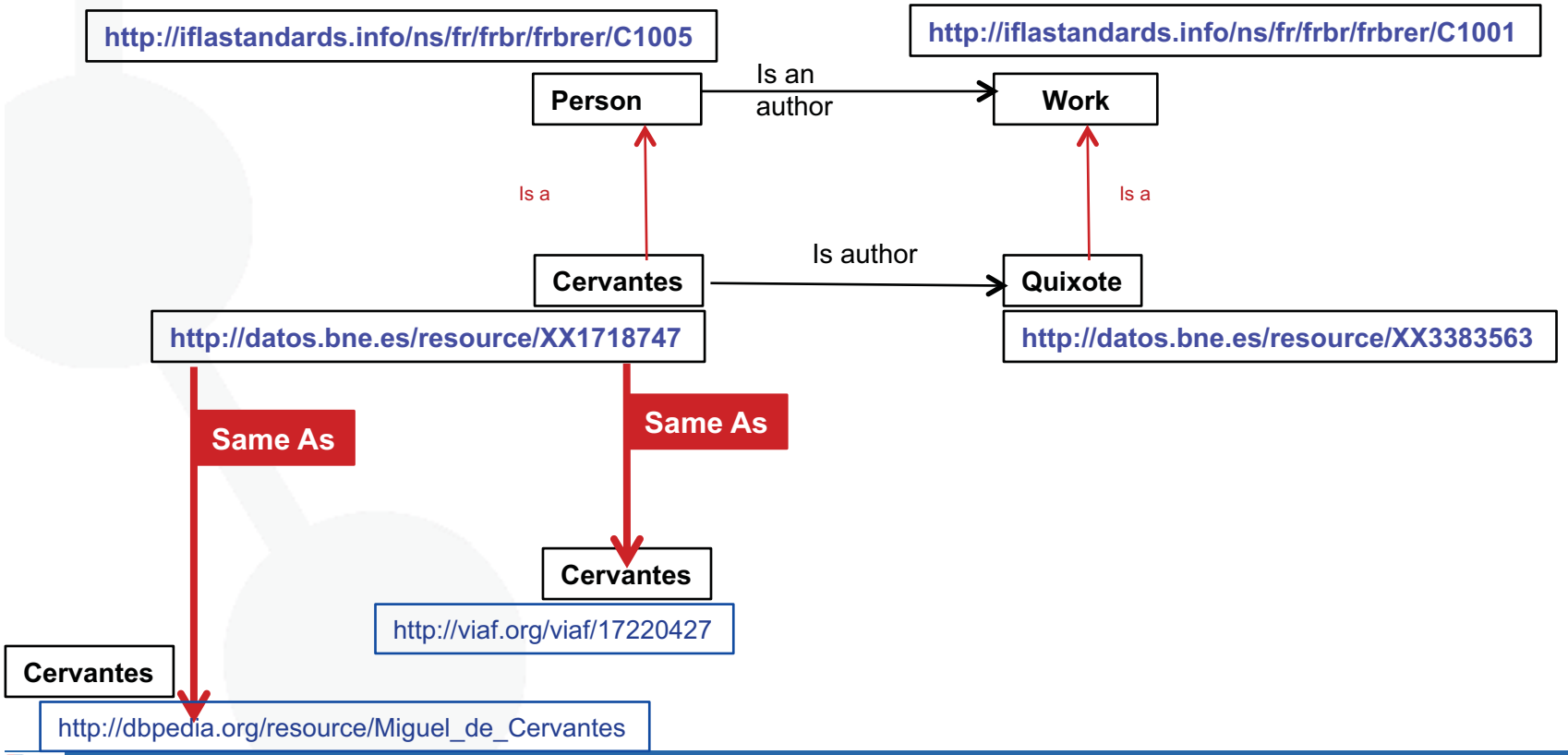
Unique Identifiers: URI

Identify a name or a resource on the Web

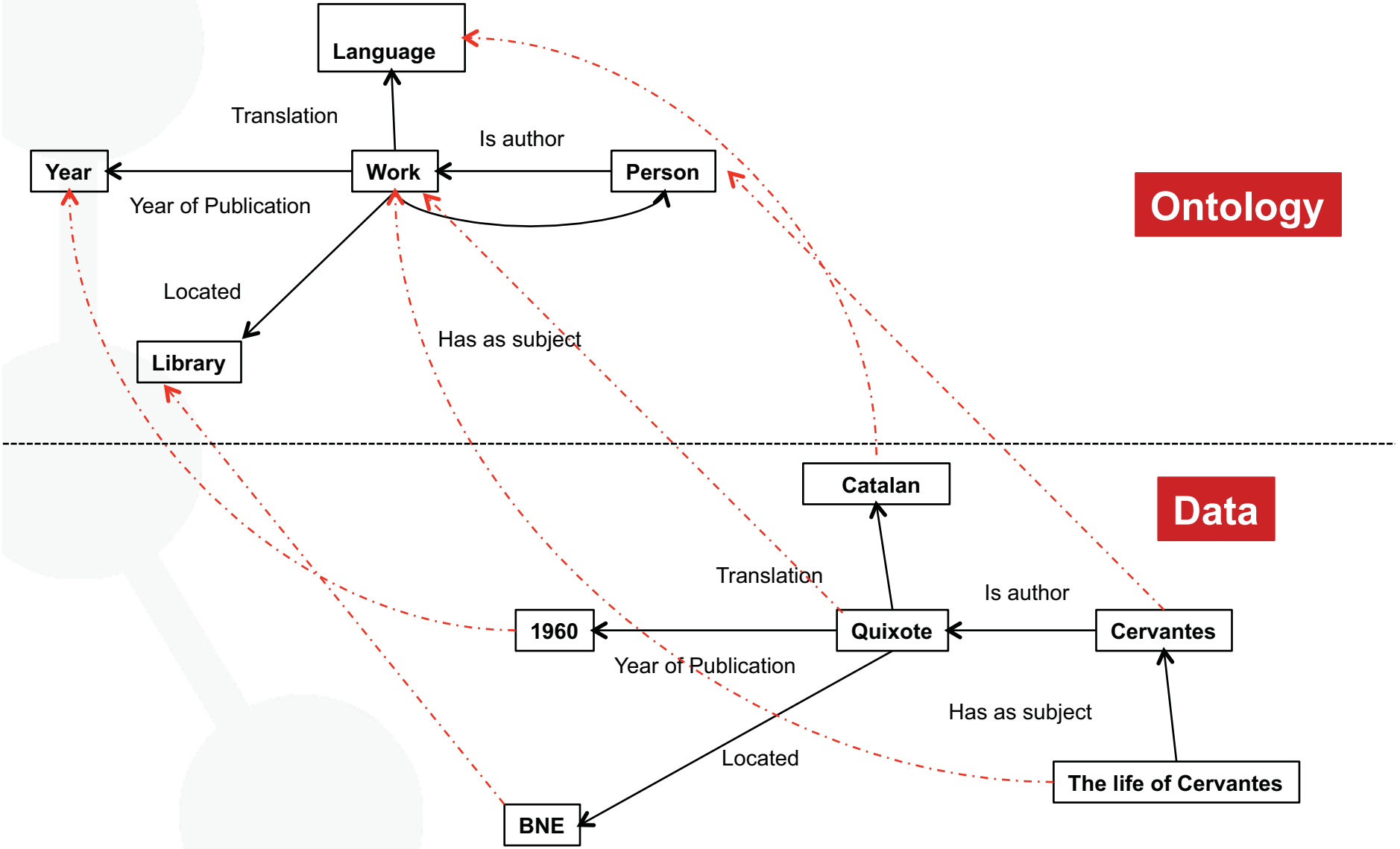
Link to other data

Same As

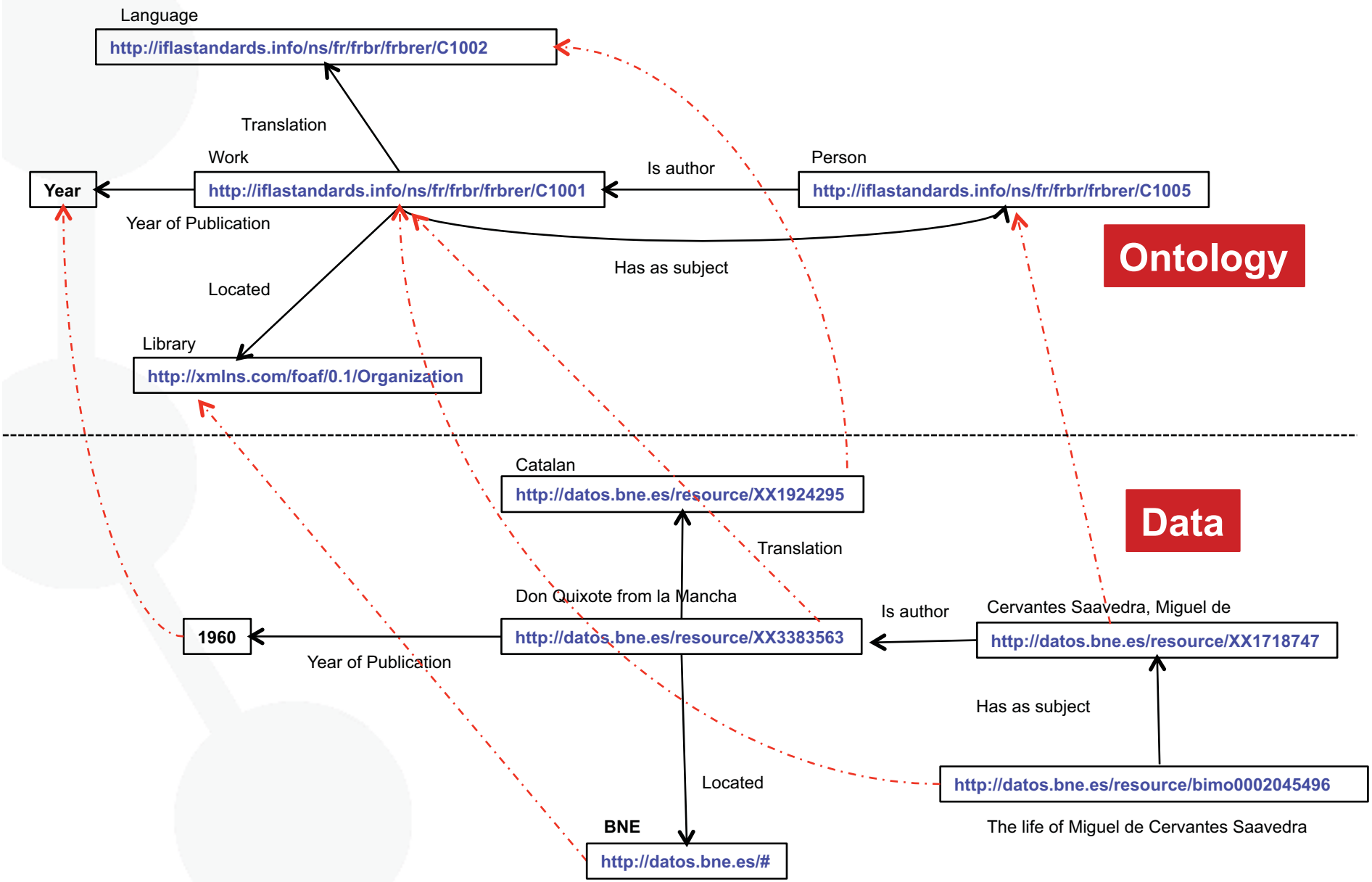
Browsing through the data



The model (Ontology) and the data



The model (Ontology) and the data



Marimba: “Data curation”

- During the successive iterations to generate RDF, **improvements in the origin records** have been produced. Some of the examples are :
  - NON - valid combinations of subfields have been identified, according to the standard MARC 21:
    - Example: 100 $a $d $1
  - Errors in the encoding of certain character strings have been identified :
    - Example: BiografÃas.
  - Errors in some of the control fields have been identified :
    - Example: A flag has been found in the field 001, which when does not correspond to the standard does not present a flag.

Marimba: Discovery of links to other datasets

- Marimba uses VIAF as a source for generating equivalence links (owl:sameAs) to other bibliographical datasets.

- For that, using a file which contains the correspondence between VIAF and the libraries participating in VIAF:
    1) Locates the IDs of the BNE and stores its corresponding VIAF.
    2) From the correponding VIAF IDs, it generates links to other libraries that also have some correspondence with these IDs.

Technological Support

- **Modelling:**
  - Open Metadata Registry
  - Neon Toolkit

- **Mapping and generation**
  - **MARiMbA: Library-oriented, supports and facilitates the entire process od transformation from MARC21 to RDF**

- **Publication:**
  - Virtuoso Universal Server
  - Pubby
  - CKAN registry
  - Sitemap4rdf

- **Exploitation:**
  - Web Applications that visualize data using SPARQL

Other data initiatives linked from libraries

- French National Library
- EU Congress Library
- German National Library
- British Library
- Spain:
  - Subject headings list for public libraries of the Ministry of Culture
    - In SKOS
    - Linked with RAMEAU and LOC subjects
  - Virtual Library of the School of Salamanca
  - W3C Use Cases:
    - Polygraph Virtual Library
    - Ontology of Cantabria´s Cultural Heritage